Korean Journal of Remote Sensing, Vol.37, No.2, 2021, pp.275~290
https://doi.org/10.7780/kjrs.2021.37.2.8

ISSN 1225-6161 (Print)
ISSN 2287-9307 (Online)

Article

기계학습을 활용한 동아시아 지역의 TROPOMI 기반 SO_2 지상농도 추정

최현영[1)] · 강유진[1)] · 임정호[2)†]

Estimation of TROPOMI-derived Ground-level SO_2 Concentrations Using Machine Learning Over East Asia

Hyunyoung Choi[1)] · Yoojin Kang[1)] · Jungho Im[2)†]

Abstract: Sulfur dioxide (SO_2) in the atmosphere is mainly generated from anthropogenic emission sources. It forms ultra-fine particulate matter through chemical reaction and has harmful effect on both the environment and human health. In particular, ground-level SO_2 concentrations are closely related to human activities. Satellite observations such as TROPOMI (TROPOspheric Monitoring Instrument)-derived column density data can provide spatially continuous monitoring of ground-level SO_2 concentrations. This study aims to propose a 2-step residual corrected model to estimate ground-level SO_2 concentrations through the synergistic use of satellite data and numerical model output. Random forest machine learning was adopted in the 2-step residual corrected model. The proposed model was evaluated through three cross-validations (i.e., random, spatial and temporal). The results showed that the model produced slopes of 1.14-1.25, R values of 0.55-0.65, and relative root-mean-square-error of 58-63%, which were improved by 10% for slopes and 3% for R and rRMSE when compared to the model without residual correction. The model performance by country was slightly reduced in Japan, often resulting in overestimation, where the sample size was small, and the concentration level was relatively low. The spatial and temporal distributions of SO_2 produced by the model agreed with those of the *in-situ* measurements, especially over Yangtze River Delta in China and Seoul Metropolitan Area in South Korea, which are highly dependent on the characteristics of anthropogenic emission sources. The model proposed in this study can be used for long-term monitoring of ground-level SO_2 concentrations on both the spatial and temporal domains.

Key Words: ground-level SO_2 concentrations, TROPOMI, machine learning, residual correction

Received April 6, 2021; Revised April 15, 2021; Accepted April 16, 2021; Published online April 22, 2021

[1)] 울산과학기술원 도시환경공학부 석·박사과정생 (Combined MS/PhD Student, School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

[2)] 울산과학기술원 도시환경공학부 정교수 (Professor, School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

[†] Corresponding Author: Jungho Im (ersgis@unist.ac.kr)

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

요약: 대기 중의 이산화황(SO_2)은 주로 인위적 배출원에 의해 발생하며 화학 반응을 통해 (초)미세먼지를 형성하여 직간접적으로 주변 환경 및 인체 건강에 해로운 영향을 주는 물질이다. 특히 지상에서의 농도는 인간 활동과 밀접한 관련이 있어 모니터링의 필요성이 매우 크다. 따라서, 본 연구에서는 TROPOMI SO_2 연직 컬럼 농도 산출물 및 타 위성 산물과 모델 산출물 등을 융합 활용하여 기계학습 기법에 적용하여 SO_2 지상 농도 추정 모델을 개발하였다. 기계학습 기법으로는 널리 활용되고 있는 RF(Random Forest)에 잔차 보정 과정을 결합한 2-step 잔차 보정 RF를 적용하였다. 개발된 모델은 무작위, 공간 및 시간별 10-fold 교차 검증을 통하여 검증하였으며, 기울기(slope) 값이 1.14-1.25, 상관계수(R) 값이 0.55-0.65, rRMSE 값이 약 58-63% 정도로 나타났다. 이는 잔차 보정이 적용되지 않은 기존의 RF 대비 slope의 경우 약 10%, R과 rRMSE의 경우 약 3% 가량 향상된 결과를 보인다. 국가별로 나누어 분석하였을 때에는 샘플 수가 적고 SO_2의 전반적인 농도가 낮은 일본 지역에서의 공간별 10-fold 교차검증 성능이 소폭 감소하는 것으로 나타났다. SO_2 지상농도 분포를 계절별로 표출하였을 때, 일본의 경우 다른 지역 대비 연중 저농도가 관찰되며 높은 결측 값 비율로 인하여 관측소 농도 대비 2-step 잔차 보정 RF 모델에서 과대 모의하는 경향이 관찰되었다. 대표적 고농도 발생지인 중국의 YRD(Yangtze River Delta) 와 한국의 SMA(Seoul Metropolitan Area)의 계절적 분포 변화를 추가적으로 분석하였을 때, 연료 연소로 인한 겨울철 농도 증가 패턴이 나타났다. 이는 인위적 배출원의 영향을 크게 받는 SO_2의 시공간적인 분포 특성을 잘 반영하고 있는 결과이다. 따라서, 본 연구를 통하여 제안한 모델은 장기적으로 SO_2 지상 농도의 시공간적 분포를 파악하는 데에 활용될 수 있을 것으로 기대된다.

1. 서론

대기 중의 이산화황(SO_2)은 주로 연료 연소 등의 인위적 배출원에 의하여 발생하며 빠르게 산화되어 황산염의 형태로 변환된다. 이는 대기 중 화학 반응을 통하여 초미세먼지를 형성하는 주요 전구 물질이다(Kharol *et al.*, 2017; Seo *et al.*, 2020; Qin *et al.*, 2020). 따라서, SO_2는 초미세먼지를 통한 간접적인 영향뿐만 아니라 식생과 건물 부식 등 주변 환경과 인체 건강에 직접적으로 해로운 영향을 미치는 물질이다(Zhan *et al.*, 2018b). 최근 들어 동아시아 지역에서는 급격한 경제 발전으로 인해서 대기 오염이 심각한 환경 문제로 인식되고 있다(Bauduin *et al.*, 2016; Shah *et al.*, 2020; Shakeel *et al.*, 2015). 게다가 SO_2는 수 시간에서 수 일 정도의 짧은 체류 시간으로 인해서 대류권 하부의 배출원 근처에 머무르는 경향이 있다(Qin *et al.*, 2020). 따라서, 대기 오염은 인간 활동과 밀접한 관련이 있는 사회경제적인 이슈이며, 지상의 대기오염 물질 농도를 모니터링하는 것의 중요성이 제기되고 있다(Baawain and Al-Serihi, 2013).

위성 자료는 넓은 지역에 대해서 공간적으로 연속적인 정보를 제공할 수 있다는 장점이 있다(Fernandes *et al.*, 2019). OMI(Ozone Monitoring Instrument)와 TROPOMI (TROPOspheric Monitoring Instrument)를 비롯한 여러 위성들은 SO_2 연직 컬럼 농도 정보를 제공한다. 특히 TROPOMI는 가장 최근에 발사되어 2018년 5월부터 기존 위성에 비해 고해상도(5.5 km × 3.5 km)로 정보를 제공할 수 있다(Huang and Sun, 2020). 다만 위성이 제공하는 정보는 연직 컬럼 농도이므로 인간 생활과 밀접한 관련이 있는 지상 농도에 대한 직접적인 정보를 제공할 수 없다는 한계점이 있다. 위성에서 관측한 연직 컬럼 농도는 지상의 대기 오염 물질 농도와 단순 선형 관계를 가지지 않는다. 특히, 위성에서는 대기 중 O_3에 의한 흡수와 에어로졸 산란에 의해서 대기 하층부에서의 SO_2 신호에 대한 불확실성이 크다(Combrink *et al.*, 1995; Fioletov *et al.*, 2011). 따라서, 연직 컬럼 농도와 지상 농도 간의 비선형 관계를 풀어내기 위하여 다양한 방법이 시도되어 왔다.

많은 연구들이 화학 수송 모델, 위성 기반의 통계적 모델, 관측소 자료를 이용한 공간적인 내삽과 같은 다양한 방법을 통하여 지상 농도의 연속적인 분포를 산출해왔다(Berman *et al.*, 2015; Young *et al.*, 2016). 모델을 활용하는 연구에서는 모델 기반의 SO_2와 위성 기반의 SO_2를 이용하여 단순 관계식을 이용해 연직 컬럼 농도를 지상 농도로 변환하는 방식이 활용되었다(Kharol *et al.*, 2017). 최근에는 지상 관측소를 실측 값으로 하여 이를 종속 변수로써 활용하는 기계학습 기법 또한 적용되고 있다. 가장 널리 이용되는 방법은 RF(Random Forest)와 Artificial Neural Network이다(Ahmad *et al.*, 2019; Huang

et al., 2018; Li *et al.*, 2020b). 모델 구축에 활용된 자료로는 위성의 SO_2 연직 컬럼 농도, 식생 지수 등의 위성 기반 산출물, 기상 자료, 고도 등의 자료들이 주로 활용되어 왔다(Li *et al.*, 2020b; Zhan *et al.*, 2018b). 기계학습 기법은 다른 지역이나 혹은 다른 날짜와 같이 훈련에 사용되지 않은 자료에 대해서 잘 작용하지 않는다는 한계점이 있다(Li *et al.*, 2020b). 그러므로, 경험적 모델이 적용될 경우 공간 및 시간적인 일반화에 대해서 검증하는 것은 매우 중요하다.

Zhan *et al.* (2018b)은 RF 기계학습 기법을 사용해서 지상의 NO_2 농도를 추정하고, 관측소가 가지는 out-of-bag(OOB) error를 이용해서 시공간적인 내삽을 통하여 한 번 더 보정하는 방법을 제안하였다. OOB error는 RF 모델에서 훈련 자료 임의 중복 추출 시 훈련에 사용되지 않은 샘플이 보여주는 에러로써 데이터의 실제 값과 예측 값 사이의 오차이다. 동일한 접근 방법으로 SO_2 지상 농도 추정에도 활용되었으며 이는 잔차를 보정함으로써 대기 오염 물질 지상 농도 추정 모델의 정확도가 향상될 수 있다는 것을 의미한다(Zhan *et al.*, 2018a). 게다가 대부분의 기계학습 기반 대기질 모니터링 연구는 실측 값 대비 추정 값이 음의 오차를 가지는 경우가 많으며(Chen *et al.*, 2019; Feng *et al.*, 2020; Li *et al.*, 2020b), 이는 잔차 보정 과정을 통하여 결과가 향상될 여지가 존재한다는 것을 의미한다. 따라서, 본 연구에서는 TROPOMI 연직 컬럼 농도 자료로부터 일별 지상 SO_2 농도 추정을 위한 기계학습 기반의 2-step 잔차 보정 모델을 개발하고자 한다. 본 연구의 목적은 TROPOMI 위성 자료를 이용하여 기존 연구 대비 고해상도(6 km × 6 km, 일별 04 UTC 타겟)의 SO_2 지상 농도를 추정하는 기계학습 기반 모델을 개발하고, 제안하는 모델의 시공간적 안정성 및 일반화 가능성을 검증하는 것이다.

2. 연구지역 및 연구자료

연구지역은 중국 동부, 한국, 일본을 포함하는 동아시아 지역(22°-48°N, 112°-148°E)으로 선정하였으며(Fig. 1), 연구 기간은 TROPOMI 위성의 가용기간을 고려하

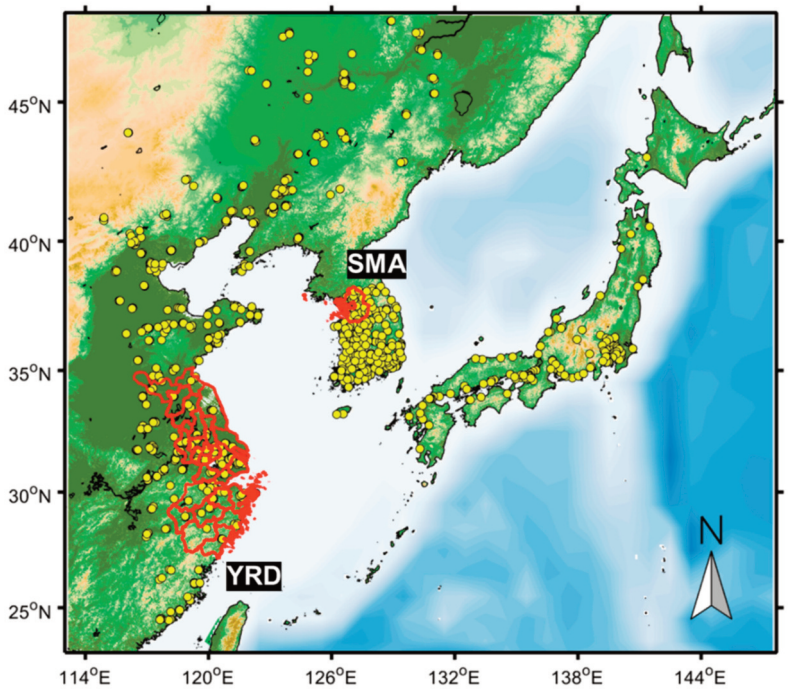

Fig. 1. The study area with SO_2 monitoring stations. The yellow points show the location of the ground monitoring stations, and the red lines represent the megacity clusters of Yangtze River Delta (YRD) and Seoul Metropolitan Area (SMA) in China and South Korea, respectively.

여 2018년 5월 14일부터 2020년 11월 12일까지로 선정하였다. 이 지역은 급속한 산업화와 도시화로 인하여 전 세계적으로 대기오염물질의 농도가 높은 지역 중 하나로 알려져 있으며, 동아시아 내에서도 상대적으로 높은 SO_2 농도 값을 보이는 두 지역을 선정하여 추가적인 시공간적 분포 분석을 진행하였다. 선정된 지역은 중국의 장강 삼각주(Yangtze River Delta; YRD)와 한국의 수도권 지역(Seoul Metropolitan Area; SMA)으로, YRD는 상하이시(Shanghai)와 장쑤성(Jiangsu), 저장성(Zhejiang) 북부를 포함하고 있으며, SMA는 한국의 서울시(Seoul), 인천시(Incheon), 경기도(Gyeonggi)를 포함하고 있다. 본 연구에서 사용된 SO_2 지상농도 관측소들의 위치와 주요 도시의 경계는 Fig. 1에 표시하였다.

본 연구에서는 연속적인 SO_2 지상농도 산출 및 분석을 위해 관측소 기반 SO_2 지상농도 자료, TROPOMI 위성자료를 포함한 다양한 위성 산출물과 수치모델 기반 기상학적 자료 등을 함께 사용하였다. SO_2 지상 관측 농도가 본 연구에서 제시하는 모델의 타겟 변수로 사용되었으며, 나머지 자료들은 입력변수로써 활용되었다. 총 32개의 변수가 SO_2 지상농도 산출 알고리즘 개발에 사용되었으며, 사용된 모든 변수들은 Table 1에 약어 및 공간 해상도와 함께 나타냈다.

SO_2 지상 관측 값의 경우, 중국 동부, 한국, 일본에서 각각 시간별 자료가 수집되었다. TROPOMI 위성의 관측 시간 (13:30 local time; 04:30 UTC)에 맞추어 04 UTC와 05 UTC 사이의 평균 농도인 04 UTC 값을 사용하였다. 중국 관측 값의 경우 베이징 시 환경 보호 모니터링 센터(Beijing Municipal Environmental Protection Monitoring Center, http://beijingair.sinaapp.com)에서 297개 측정소 자료를 오픈 API를 통해 다운받았으며, 남한의 경우 에어코리아(Airkorea, https://www.airkorea.or.kr/web)에서 제공하는 209개 측정소의 확정자료를 사용하였다. 일본의 경우 일본 국립환경과학원(National Institute for Environmental Studies, Japan, http://www.nies.go.jp)과 대기환경 지역관측시스템(Atmospheric Environmental Regional Observation System, http://soramame.taiki.go.jp/)의 223개 관측소 자료를 사용하였다. 국가별로 제공되는 SO_2 지상농도의 단위가 다르기 때문에 본 연구에서는 parts per billion(ppb)으로 통일하였으며, 이때 한국의 경우만 parts per million(ppm) 단위로 자료가 제공되어 단

Table 1. Summary of input variables used in the machine learning models in this study

Category	Variable full name	Abbr.	Resolution (km)	Category	Variable full name	Abbr.	Resolution (km)
Satellite-based data	TROPOMI SO_2 vertical column density	SO_2 VCD	5.5 × 3.5	Model-based data	Minimum temperature	Tmin	12
	GOCI Aerosol Optical Depth	AOD	6		Frictional velocity	FrictionalVelocity	
	SRTM Digital Elevation Model	DEM	0.9		Convective available potential energy	PotentialEnergy	
	MODIS Cropland area ratio	LCcrop	0.5		Surface roughness	SurfaceRoughness	
	MODIS Forest area ratio	LCforest			Latent heat net flux	LatentHeatFlux	
	MODIS Urban area ratio	LCurban			Specific humidity	SpecificHumidity	
Model-based data	Temperature	Temp	12		Accumulated maximum wind speed (1,3,5,7 days)	1D MaxWS	
	Dew-point temperature	Dew				3D MaxWS	
	Relative humidity	RH				5D MaxWS	
	Surface pressure	Psrf				7D MaxWS	
	Max wind speed	MaxWS			Wind speed	WS	
	Planetary Boundary Layer Height	PBLH			Cosine value of wind direction	Wcos	
	Visibility height above ground	Visibility			Sine value of wind direction	Wsin	
	Accumulated precipitation 3 h	AP3h		Auxiliary data	Converted day of year	DOY	–
	Temperature surface	Tsrf			Road density	RoadDens	8
	Maximum temperature	Tmax			Population density	PopDens	1

위 변환 시 다른 국가에 비해 작은 유효숫자를 가진다.

입력변수로는 다중 위성자료 및 수치모델 자료 등을 활용하여 동아시아 지역의 SO_2 지상농도 추정 알고리즘을 개발하였다. TROPOMI 센서는 Sentinel-5P 위성에 탑재되어 2017년 10월에 발사된 후 현재까지 미량기체들과 에어로졸 특성 정보를 산출하고 있으며, 본 연구에서는 일별로 전구영역에 대해 제공되는 SO_2 연직 컬럼 농도 산출물인 S5P_L2__SO2____HiR 1를 활용하였다(Veefkind *et al*., 2012; https://mirador.gsfc.nasa.gov/). SO_2는 대기 중에서 화학 반응을 통하여 (초)미세먼지를 형성하는 전구 물질로 작용하기 때문에, 에어로졸에 의한 태양 복사의 감소를 수치화한 값으로 (초)미세먼지와 밀접한 관련이 있는 변수인 에어로졸 광학 두께(Aerosol Optical Depth; AOD)를 함께 사용하였다. 본 연구에서는 대한민국 통신해양기상위성(Communication, Ocean and Meteorological Satellite; COMS), 즉 천리안 위성의 해양 탑재체인 Geostationary Ocean Color Imager(GOCI)에서 연세 에어로졸 산출(Yonsei AErosol Retrieval; YAER) 알고리즘을 통해 추출한 6 km × 6 km 해상도의 AOD 산출물을 사용하였다(Choi *et al*., 2018). 더불어 지표의 특성을 고려해주기 위해 Shuttle Radar Topography Mission (SRTM)의 수치 표고 모델(Digital Elevation Model; DEM) 산출물(https://earthexplorer.usgs.gov/)과 MODerate resolution Imaging Spectroradiometer(MODIS)의 토지피복 산출물(MCD12Q1(Sulla-Menashe and Friedl, 2018); https://search.earthdata.nasa.gov/)을 활용하였다. MCD12Q1은 500 m 해상도로 연간 토지 피복 분류(classification) 결과를 제공하며, 본 연구에서는 13 × 13 개의 인접 픽셀을 이용한 이동 창(moving window)를 활용하여 비율(ratio) 값으로 변환하여 사용하였다.

대기오염물질 농도는 기상 등 외부 환경의 영향을 받기 때문에 수치모델 기반 기상학적 변수들을 함께 사용하였다. 수치모델 자료로는 대한민국 기상청 기상자료개방포털에서 제공하는 지역예보모델(Regional Data Assimilation and Prediction System; RDAPS)을 사용하였다. 연직으로 약 80 km까지 70 층으로 구성된 모델로, 3시간 간격으로 전지구예보모델로부터 경계장을 제공받아 12 km의 공간해상도로 1일 4회(00, 06, 12, 18 UTC) 제공하는 분석장을 사용하였다. 분석장을 제공하지 않는 시간에 대해서는 시간적 내삽을 통해 자료를 생성하였다.

위성 및 수치모델 기반 자료 외에 SO_2의 인위적 배출의 시공간적 변동성 정보를 주기 위하여 기타 보조 변수로 DOY(day of year), 도로 및 인구 밀도 자료를 사용하였다. 본 연구에서 DOY는 계절을 고려하여 사인(sine) 함수를 사용해 -1에서 1 범위의 값으로 변환하여 사용하였다. 도로 밀도 자료는 GLOBIO (http://www.globio.info/download-grip-dataset)에서, 인구 밀도 자료는 NASA Socioeconomic Data and Applications Center (SEDAC, https://sedac.ciesin.columbia.edu/)에서 제공받았다.

3. 연구방법

본 연구가 제시하는 흐름은 모델 개발 및 모델 검증을 포함한 두 개의 파트로 구분할 수 있으며, Fig. 2에 나타나 있다. 본 연구에서는 SO_2 지상 농도 추정을 위하여 TROPOMI를 비롯한 위성 자료 및 모델 자료를 융합 활용하여 RF를 이용한 기계학습 기반의 모델을 개발하였다. 각 입력 변수의 공간해상도가 모두 다르기 때문에 정지궤도 위성인 GOCI 격자에 맞추어 6 km 크기로 리샘플링(resampling)하여 통일시켜주었다. 이때 이중선형보간법(bilinear interpolation)을 활용하였으며, 정지궤도 위성의 격자를 사용함으로써 고정 영역에 대한 일별 결과를 산출하였다. 기계학습 기법으로는 기존의 단일 RF 모델이 가지는 잔차를 보정하여 향상된 결과를 얻고자 잔차를 추정하는 또 다른 RF 모델을 개발하여 초기 예측 값과 잔차 예측 값을 합하여 최종 결과를 산출하는 2-step 잔차 보정 RF 모델이 개발되었다. 개발된 모델은 시간 및 공간적인 측면에서 검증하였으며 잔차 보정 과정의 효과 또한 분석되었다. 개발된 모델을 이용하여 전체 기간 동안의 연구지역 내의 SO_2 지상농도 공간 분포와 고농도 사례 관측 지역에서의 계절 평균 농도 변동 패턴 분석이 이루어졌으며, 세부 내용은 아래와 같다.

1) 모델 구축: 2–step 잔차 보정 Random Forest

RF는 훈련 시 이용되는 자료와 독립 변수를 무작위로 추출하는 다수의 의사 결정 나무를 기반으로 하는 앙

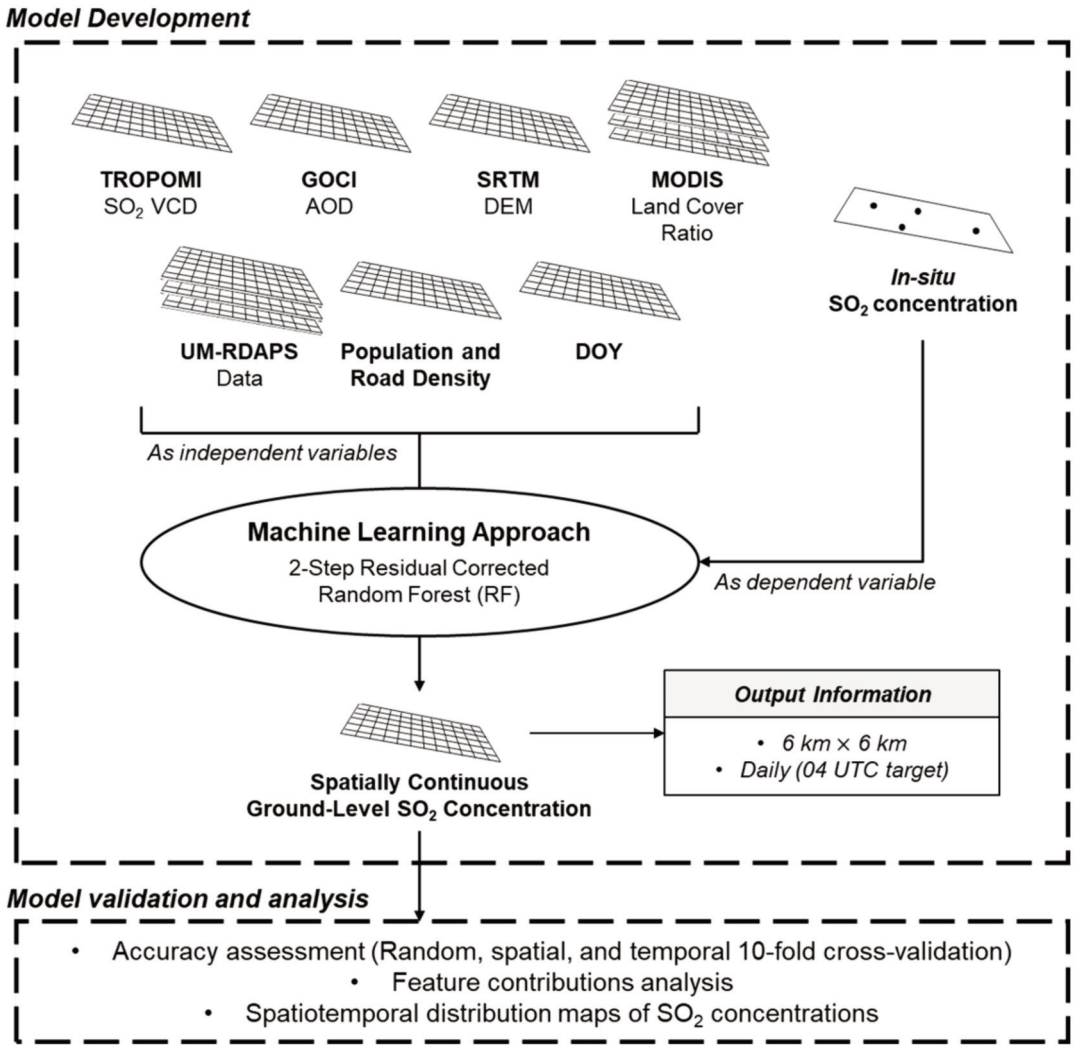

Fig. 2. Process flow diagram for estimation of ground-level SO_2 concentrations proposed in this study.

상블 학습 기법이다(Breiman, 2001). RF는 다수의 의사 결정 나무에서 나온 결과들을 평균하는 방식으로 회귀 문제의 최종 결과를 산출한다. RF가 가지는 가장 주요한 특징은 의사 결정 나무 사이의 독립성을 보장하기 위한 무작위성이다. Bootstrap과 aggregating(bagging) 방식을 통하여 훈련 자료에서 일부만을 무작위로 추출하여 사용하고, 독립 변수들 에서도 일부만을 무작위로 추출하여 의사 결정 나무를 구성한다. 이 때, 훈련에 사용되지 않은 나머지 자료를 OOB 자료라고 일컬으며 RF 모델 내부적으로 OOB error를 계산하는 검증 과정을 거친다. RF의 무작위성과 앙상블 구조는 모델의 분산을 줄이고 과적합을 방지할 수 있는 효과적인 방법이다 (Breiman, 2001; Liaw and Wiener, 2002).

RF는 bagging 과정을 통하여 분산을 줄일 수 있으나, 편차에 대해서는 큰 변화가 발생하지 않는다(Zhang and Lu, 2012). 그러므로 잔차 보정 과정을 거치는 RF는 향상된 결과를 산출할 수 있다. 모델의 잔차를 보정하기 위해서 RF 초기 결과 값에 단순 상수를 더해주는 방법, OOB 자료를 이용해서 예측 값과 실측 값 사이의 선형 관계를 세워 보정해주는 방법, 혹은 기계학습을 한 번 더 사용하여 잔차 예측 모델을 구축한 후 이를 초기 결과에 더해주는 방법 등 다양한 방법이 사용되었다(Song, 2015). Song (2015)의 연구 결과에 의하면 잔차 예측 모델을 따로 구축한 후 초기 결과에 더해주는 방법이 가

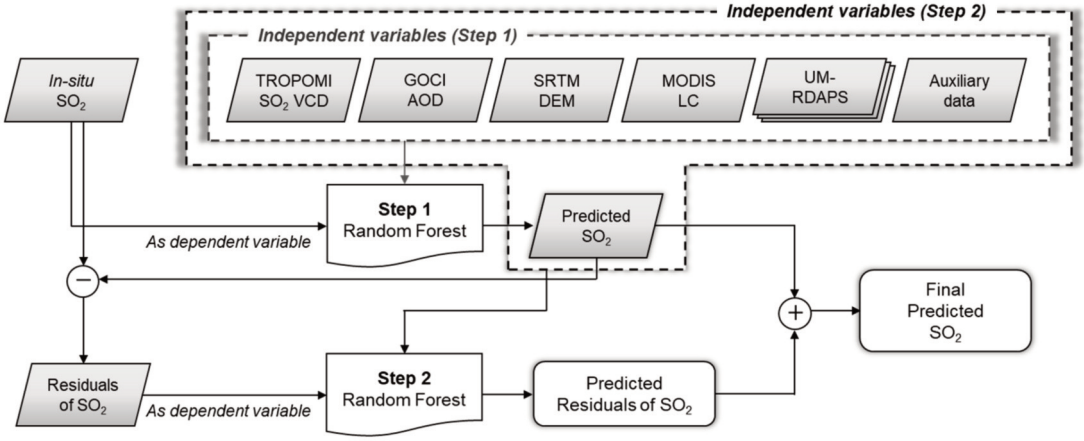

Fig. 3. Process flow diagram of the 2-step residual corrected random forest model.

장 좋은 성능을 보여주었으므로 본 연구에서는 해당 방법을 통하여 2-step 잔차 보정 RF 모델을 개발하였다.

2-step 잔차 보정 RF 모델의 step 1은 SO_2 지상 농도를 추정하는 과정이며, 32개의 독립 변수들을 이용하여 RF 기계학습에 적용하여 개발되었다. 이후 step 1에서 산출되는 예측 값과 실측 값 사이의 잔차를 계산하여 step 2에서 이용될 종속 변수를 구축하였다. Step 2에서는 앞서 계산된 잔차를 예측하는 RF 모델을 개발하였으며, 이 과정에서 사용되는 독립 변수는 step 1에서 사용된 32개의 변수 외에 step 1에서 산출된 SO_2 지상 농도 예측값을 함께 활용하므로 총 33 개의 독립 변수를 이용하였다. 최종 결과는 step 1 RF 모델의 SO_2 지상 농도 예측값과 step 2 RF 모델의 잔차 예측 값을 더하여 산출된다 (Fig. 3).

2) 모델 평가

개발된 모델을 평가하기 위하여 기울기(slope), 상관계수(R), root-mean-square-error(RMSE), relative RMSE (rRMSE), mean biased error(MBE), percent change(PC)의 6개의 통계 지표가 사용되었다. 각 통계 지표를 계산하는 수식은 수식 (1)-(6)에 나타냈다.

$$slope = \frac{\Delta y}{\Delta x} \quad (1)$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)}{n}} \quad (3)$$

$$rRMSE = \frac{RMSE}{\bar{y}} \times 100 \quad (4)$$

$$MBE = \frac{\sum_{i=1}^{n}(x_i - y_i)}{n} \quad (5)$$

$$PC = 100 \times \frac{(V_N - V_O)}{V_O} \quad (6)$$

수식에서 n은 전체 자료의 개수, x_i와 y_i는 각각 모델 예측 값과 실측 값을 의미하며, $\bar{x}$와 $\bar{y}$는 모델 예측 값과 실측 값들의 평균값을 의미한다. PC 계산에 활용되는 V_N과 V_O는 각각 앞서 계산된 다섯 개의 통계 지표들의 step 2의 결과값과 step 1의 결과값을 의미한다. PC는 다른 다섯 개의 통계 지표에 대해서 step 1의 결과에 비하여 step 2에서 향상된 정도를 평가하기 위한 지표로 사용되었다.

본 연구에서는 개발된 모델의 시공간적인 안정성을 검증하기 위하여 무작위 교차 검증(Random cross-validation; RDCV), 공간적인 교차 검증(Spatial cross-validation; SPCV), 시간적인 교차 검증(Temporal cross-validation; TPCV) 세 가지의 10-fold 교차 검증 방법을 통하여 모델의 성능을 평가하였다. 10-fold 교차 검증은 연구에 사용되는 전체 자료를 동일한 개수로 10개의 서브그룹으로 구분하여 나눈 뒤, 9개의 서브그룹 자료를 이용하여 모델을 훈련하고 나머지 1개의 서브그룹 샘플에 대하여 검증에 이용하는 방식을 10번 반복하여 최종 결과를 도출해내는 검증 방법이다. 무작위 교차 검증은 개발된 모델의 전반적인 성능 평가를 위하여 사용되며, 개발된 모델이 훈련에 사용되지 않은 위치 및 시간에서 가질 수 있는 성

능 평가를 위하여 추가적으로 공간적인 교차 검증과 시간적인 교차 검증을 실시하였다. 공간적인 교차 검증은 관측소를 기준으로 10개의 서브그룹으로 구분하였으며, 시간적인 교차 검증은 DOY를 기준으로 10개의 서브그룹으로 구분하였다.

4. 연구결과 및 토의

1) 모델 검증

Table 2는 2-step 잔차 보정 RF 모델의 10-fold 교차검증 결과와 단일 RF 모델 결과 대비 PC(%)를 나타낸다. 본 연구에서 제안된 2-step 잔차 보정 RF 모델 결과 세 가지 검증(RDCV, SPCV, TPCV) 측면에서 slope=1.14-1.25, R=0.55-0.65, MBE=0.0238-0.0274 ppb 수준으로 지상 SO_2 농도를 추정하고 있는 것을 확인할 수 있다. 다섯 가지 평가 지표들을 통해 살펴보았을 때, 2-step 잔차 보정 RF 모델을 적용했을 때 세 가지 검증 측면에서 모두 그 성능이 개선된 것을 확인할 수 있다. 특히, 단일 RF 대비 2-step 잔차 보정 RF 모델을 적용했을 때 slope와 MBE 측면에서 큰 성능 개선이 있었다. Slope의 경우 대략 9-10% 정도 감소하며 일대일 선에 가까워지는 양상을 보였으며, MBE의 경우 대략 76-80% 가량 눈에 띄게 감소하는 경향을 보였다. 이 외에도 모든 지표 측면에서의 PC를 확인해보았을 때, 2-step 잔차 보정 RF 모델이 단일 RF에서 발생하는 잔차를 보정해줌으로써 전반적인 성능 향상을 이끌어 냈음을 의미한다. 또한, 구축된 모델을 여러 측면에서 검증하기 위해 세 가지 검증을 적용한 결과, RDCV의 경우 동일 관측소와 DOY의 자료가 모두 분리되지 않고 랜덤하게 섞여 있어 가장 높은 검증 정확도를 보였다. 이는 RDCV의 경우 같은 날짜 혹은 같은 관측소의 샘플이 일부는 훈련 서브셋에, 나머지는 검증 서브셋에 포함될 수 있기 때문이다. 이에 비해 SPCV와 TPCV는 각각 관측소와 DOY를 기준으로 분리된 검증 서브그룹으로 구성되어 있기 때문에 검증 결과가 상대적으로 떨어지는 것을 확인할 수 있다. 기계학습 기반의 모델이 훈련에 포함되지 않은 공간 혹은 시간에 대하여 정확성이 감소하는 경향은 흔히 관찰되는 특성이며(Huang *et al.*, 2018; Choi *et al.*, 2020), 전반적으로 성능이 소폭 감소하기는 하지만 slope 측면에서는 오히려 좋은 결과를 보이며 훈련되지 않은 시공간 자료에 대해서도 어느정도 안정성을 가지는 모델임을 보여준다. 또한, 기존의 많은 SO_2 지상농도 산출 연구들이 월간 혹은 연간 평균을 통해 15 km 이상의 낮은 시공간 해상도로 결과를 산출하는 것을 고려하면, 본 연구 결과는 일별(04 UTC 타겟) 6 km 해상도로 결과를 제공하며 기존 연구와 비슷한 성능을 보이므로 본 지표들이 상당히 의미 있는 값으로 판단된다(Kharol *et al.*, 2017; Zhang *et al.*, 2018; Li *et al.*, 2020a).

Table 3은 본 연구에서 제시한 모델의 국가별 정확도 검증을 위해 관측소를 기준으로 분리된 SPCV 결과를 활용하여 중국, 한국, 일본에 대한 각각의 성능을 검증

Table 2. Accuracy statistics from the three 10-fold cross-validations of the 2-step residual corrected random forest model and the improvement from the RF model without residual correction. The degree of improvement of each indicator is calculated in percent change (%)

		RDCV	SPCV	TPCV
2-Step Residual Corrected RF	Slope	1.25	1.14	1.19
	R	0.65	0.55	0.56
	RMSE (ppb)	2.5711	2.7976	2.7810
	rRMSE (%)	57.8	62.9	62.5
	MBE (ppb)	0.0259	0.0238	0.0274
		RDCV	SPCV	TPCV
PC (%)	Slope	-10.07	-9.52	-9.16
	R	3.17	1.85	1.82
	RMSE	-3.21	-1.57	-1.79
	rRMSE	-3.18	-1.56	-1.73
	MBE	-76.94	-79.68	-76.50

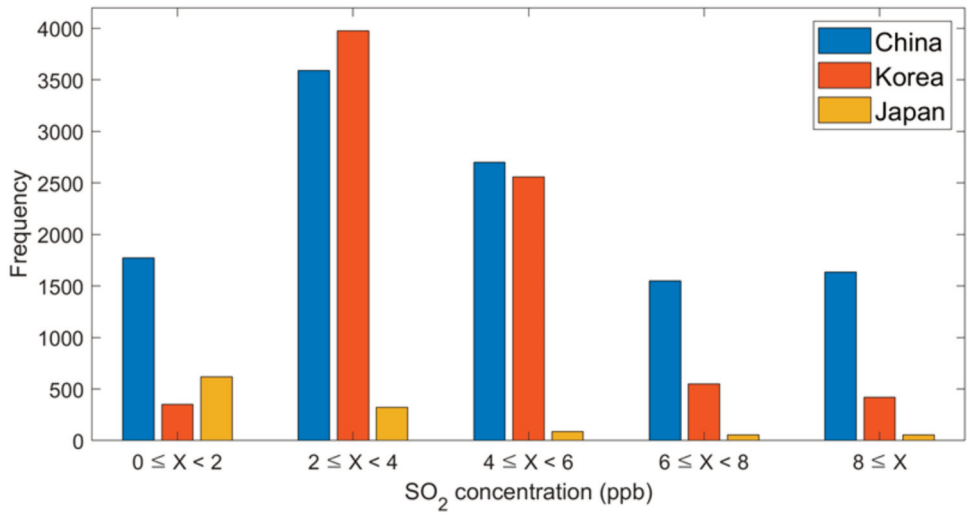

Fig. 4. The distribution of the station-based SO_2 concentrations of the training samples by country.

Table 3. Accuracy statistics from the spatial cross-validation of the 2-step residual corrected random forest model by country

	China	South Korea	Japan
Slope	1.16	1.15	0.63
R	0.57	0.45	0.25
RMSE (ppb)	3.01	2.39	3.17
rRMSE (%)	60.5	61.7	87.3
MBE (ppb)	0.73	0.52	-0.12

한 결과이다. 동일하게 다섯가지 지표를 통해 평가하였을 때 중국, 한국, 일본 순으로 우세한 검증 정확도를 보였으며, 중국과 한국에 비해 일본의 경우 눈에 띄게 감소하는 정확도를 보였다. 이러한 국가별 정확도 차이는 훈련자료의 지상 관측 SO_2 농도 값의 분포와 샘플 수의 차이에서 기인하는 결과로, 일본은 전체 연구기간 동안 중국과 한국에 비해 저농도 샘플의 비율이 매우 높으며 (Fig. 4), 위성자료의 구름에 의한 결측 등으로 훈련 샘플의 수가 상대적으로 매우 적다. 전체 20,227개의 훈련 샘플 중 중국 관측소 자료는 11,242개(55%), 한국 7,851개(39%), 일본 1,134개(6%)로 중국과 한국에 편향된 자료임을 확인할 수 있다. 훈련자료의 중국, 한국, 일본의 평균 지상관측 SO_2 농도는 각각 5.02, 3.94, 2.41 ppb로 평균 농도 역시 일본에서 매우 낮게 나타난다. 일본에서의 낮은 농도 분포와 부족한 훈련 샘플의 수에 의해 SPCV를 통한 검증 시 해당 관측소에 대한 정보가 훈련에 사용되지 않았기 때문에 국가 별 검증 정확도 차이가 크게 나는 것을 확인할 수 있다. 평가 지표 측면에서 살펴보면, 중국과 한국의 경우 slope가 각각 1.16과 1.15로 상대적으로 과소 모의하는 결과를 보여주는 반면, 일본의 경우 slope 0.63으로 저농도 구간에 대하여 과대 모의하고 있는 결과를 보인다(Table 3). R과 rRMSE 측면에서도 일본에서 저하된 성능이 관찰되며, RMSE나 MBE는 일본의 낮은 농도 범위에 의해 크게 차이가 없거나 더 나은 값을 보였다. 이러한 샘플 분포의 불균형에 의해 발생하는 지역 간 정확도 차이는 추후 부족한 농도 구간에 대하여 오버샘플링(oversampling) 기법을 적용해 샘플의 분포를 보완해줄 경우 모델 개선 가능성이 기대된다.

2) 변수 중요도

Fig. 5는 본 연구에서 제안된 2-step 잔차 보정 알고리즘의 step 1과 step 2 RF 모델의 상대적인 변수 중요도를 나타낸 것이다. 변수 중요도는 R studio 프로그램의 ranger package에서 제공하는 impurity를 이용하여 계산하였다. RF는 다수의 의사 결정 나무로부터 평균 예측치를 출력하며, 각 의사 결정 나무는 각 영역의 불순도(impurity)가 감소하는 방향으로 가지를 뻗어 나가며 모델을 구축한다. 회귀 문제의 경우에 impurity는 예측 값에 대한 분산을 의미한다(Wright *et al.*, 2018). 즉, 본 연구에서는 해당하는 변수가 불순도를 어떻게 변화시키는지를 이용하여 계산한 변수 중요도를 사용하였다. SO_2

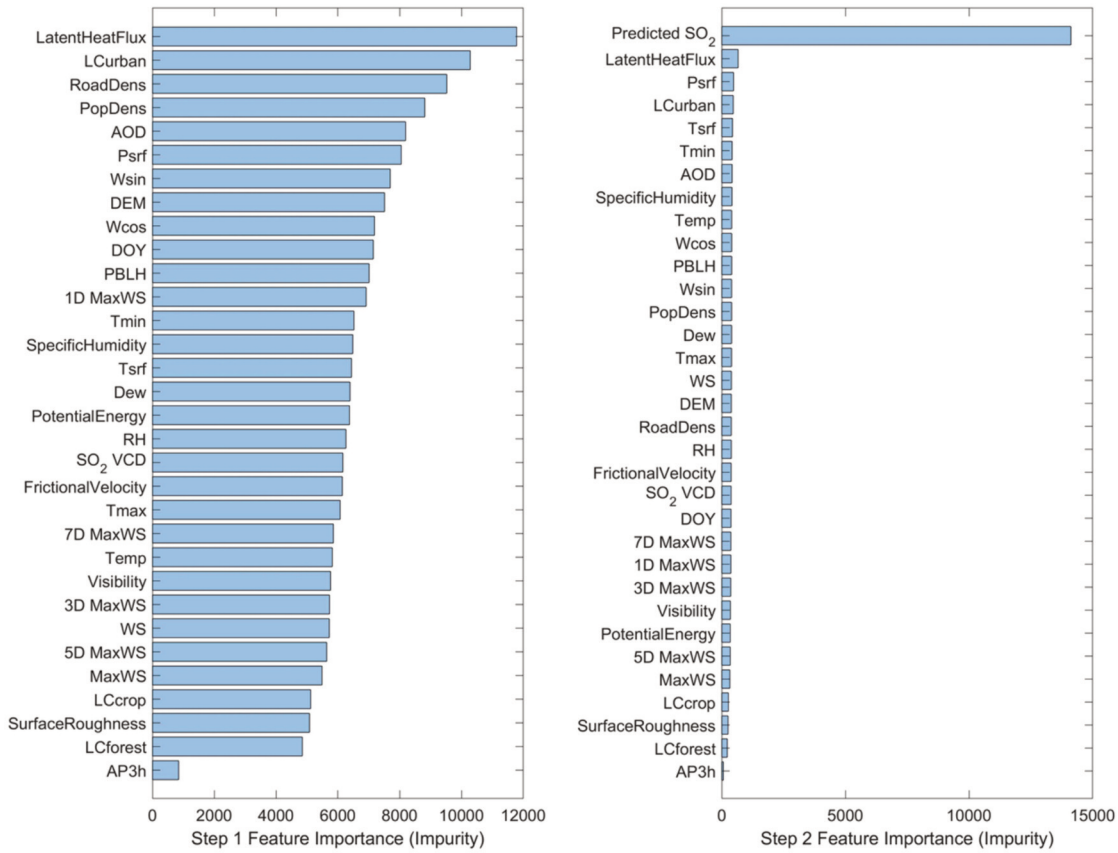

Fig. 5. Relative variable importance of step 1 (left) and step 2 (right) by random forest.

지상농도 추정 모델에서 기여도가 높은 상위 네 개의 변수는 LatentHeatFlux, LCurban, RoadDens, PopDens이다. 이 중에서 LatentHeatFlux를 제외한 세개의 변수는 모두 도시화와 인위적 배출 요인을 고려해 줄 수 있는 변수로써, 주로 인위적 요인에 의하여 발생하는 SO_2 지상 농도의 분포를 잘 반영해주기 때문에 기여도가 높게 나타난 것으로 판단된다. LatentHeatFlux의 경우 지상 SO_2 농도와의 상호 작용이나 인과 관계에 대해서 명백히 밝혀진 바는 없지만, LatentHeatFlux는 위도와 계절에 따라서 큰 변동이 있다(Zhang and Rossow, 1997). 저위도에서 높은 경향이 관찰되며, 육상에서 여름철의 LatentHeatFlux는 높고 반대로 겨울철에는 낮아진다. 따라서, 이러한 공간적인 특성과 계절성이 SO_2 지상 농도 분포에 영향을 준 것으로 추측된다. SO_2 VCD의 변수 중요도가 낮게 나타난 까닭은 대류권이 아닌 전층 컬럼 농도이기 때문이다. 전층 컬럼 농도는 에어로졸 산란과 오존 흡수에 의해 쉽게 영향을 받기 때문에 낮은 대기층에서의 SO_2 정보를 반영하기 어렵다는 한계점이 있다(Fioletov *et al.*, 2011). 실제 SO_2 VCD와 지상 SO_2 농도 간의 상관관계는 0.1 수준에 그치는 낮은 값을 보이는 것으로 확인되었다.

Step 2에서는 step 1에서 추정된 지상 SO_2 농도 값이 압도적으로 높은 기여도를 보였다. 즉, step 1에서 추정된 지상 SO_2 농도 값이 잔차 값의 대부분을 결정한다는 것이다. Predicted SO_2는 step 2 RF의 타겟인 잔차 값과 0.76 상관관계를 보였다. 즉, step 1에서 predicted SO_2 농도의 값이 클수록 더 큰 잔차를 보정하도록 모델이 구축되었음을 추측할 수 있다.

3) SO_2의 시공간적 분포 분석

Fig. 6는 전체 연구기간 동안의 평균 SO_2 지상농도의 공간적 분포와 각 픽셀 별 전체 기간 중 결과 값이 존재하는 날짜의 비율을 나타낸 것이다. 본 연구에서는 입력자료에 여러 위성의 산출물이 함께 사용되었기 때문

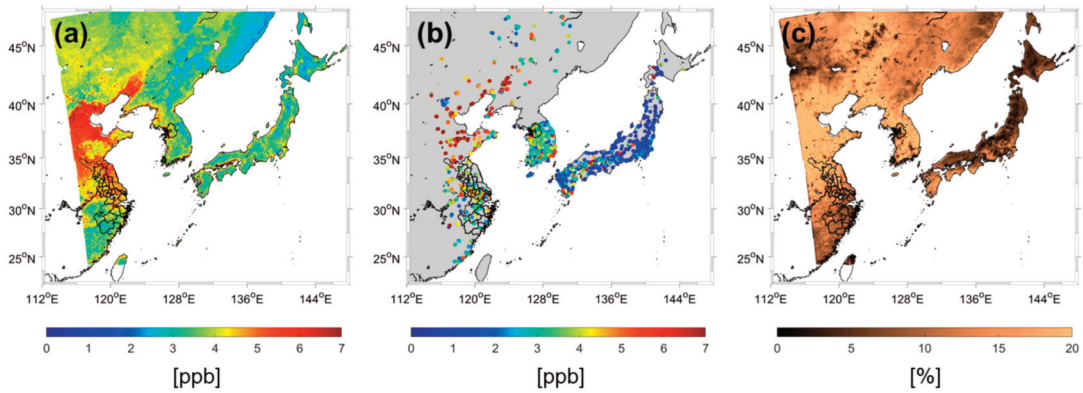

Fig. 6. Spatial distribution of the mean SO_2 ground-level concentrations during the whole study period (June 2018 – October 2020): (a) 2-step residual corrected random forest model predictions, (b) *in-situ* observations, (c) temporal data coverage (%). The unit of SO_2 concentration is part per billion (ppb).

에 구름이나 높은 알베도 등의 요인에 의해 결측 값이 생길 경우 결과를 산출할 수 없다는 한계를 가진다. 따라서, 본 연구에서 제시하고 있는 모델의 결과를 산출할 때에, 각 픽셀에 대한 신뢰도 정보 및 대표성 여부를 판단하기 위해 전체 기간 중 결과값이 산출되는 날짜의 비율을 함께 나타냄으로써 결과를 해석하는 데 보조 자료로 활용하였다. Fig. 6의 모델 예측 값과 관측 값을 비교하였을 때 고농도가 관찰되는 지역인 중국 북동부와 남한의 수도권 지역 및 부산, 울산 등의 지역에서 공간적 분포가 잘 일치하는 것으로 나타났다. 일본의 경우 관측소 농도 대비 2-step 잔차 보정 모델에서 과대 모의하고 있는 것을 확인할 수 있는데, 이는 위의 모델 검증 부분에서도 언급했듯이 동아시아 영역의 다른 지역 대비 일본의 SO_2 지상농도 관측 값이 낮은 값 분포를 가지며 결측 값의 비율이 커 훈련에 불균형이 발생한 것에서 기인한다. Fig. 6(c)의 공간 분포에서 역시 확인할 수 있듯이 일본 전역과 YRD 지역, 내몽고자치구 지역에서 약 10% 미만의 낮은 결과 표출 비율을 보이며 불확실성을 내포하고 있다. 이러한 낮은 산출 빈도는 일부 위성자료(TROPOMI SO_2 VCD와 GOCI AOD)의 결측 값에서 기인하지만, 해당 변수들은 높은 변수 기여도를 보이거나(Fig. 5) SO_2 컬럼 정보를 제공하는 변수로써 본 연구에서 중요한 역할을 하는 것을 확인할 수 있다. 최근 이러한 위성 자료의 한계를 극복하기 위해 기계학습 혹은 딥러닝(deep learning)을 활용한 다양한 보간(interpolation) 기법 등에 관한 연구가 진행되고 있으며(Wang and Sun, 2019; Yuan *et al.*, 2020; Xiao *et al.*, 2021), 추후 이를 통한 입력자료 보완 시 모델 성능의 향상을 기대할 수 있다.

SO_2는 1차 대기오염물질로 인간 활동과 관련된 인위적 배출의 영향을 크게 받으며 Fig. 6와 같이 뚜렷한 지역적 분포 특성을 보인다. 따라서, 교통이나 산업 및 주거 시설에 의해 고농도 SO_2가 빈번히 발생하는 지역에 대하여 모델의 시공간적 모의 능력을 확인하기 위하여 Fig. 1에서 선정한 두 지역인 YRD와 SMA에 대하여 모델 추정 값과 관측 값의 계절적 분포 변화를 추가적으로 분석하였다. Fig. 7와 Fig. 8은 각각 YRD와 SMA 지역의 계절별 SO_2 지상농도 분포를 나타내며 Fig. 6에서와 동일하게 결과 값이 존재하는 날짜의 비율(%)을 함께 나타냈다. 이때 3월부터 5월까지는 봄, 6월부터 8월까지는 여름, 9월부터 11월까지는 가을, 12월부터 2월까지는 겨울로 구분하였다. Fig. 7와 Fig. 8은 모두 겨울철에 농도가 증가하고 여름철에 감소하는 SO_2의 일반적 계절 변동 특성을 잘 모의하고 있다. 동아시아 SO_2의 이러한 계절 변동 패턴이 나타나는 까닭은 겨울철 난방시스템과 안정적인 대기환경이 복합적으로 작용해 고농도가 관측되기 쉬운 조건을 형성하는 반면 여름철에는 상대적으로 적은 인위적 배출과 강수에 의한 탁월한 세정효과, SO_2가 광화학적 반응에 의해 황산염(sulfate)으로 변환되는 과정이 활발하게 일어나기 때문이다(Lin *et al.*, 2019; Choi *et al.*, 2020).

Fig. 7은 중국의 YRD 지역에서의 계절별 SO_2 농도 변화를 잘 반영하고 있다. 관측소 값과 모델 추정 값은 대부분 일치하는 경향을 보였으며, 두 값 사이의 오차가

큰 사례는 대부분 Fig. 7(b)에서 결과값 산출 가능 날짜의 비율이 적은 픽셀에서 발생하는 것을 확인할 수 있다. 특히 이 지역에서 여름철의 결과 산출 비율은 5% 이내로, 잦은 강수와 구름에 의한 결측으로 산출 대표성이 떨어진다. 이러한 까닭으로 인해 여름철 YRD 지역에서 관측 값 대비 모델 추정 값이 상대적으로 과대 모의하고 있지만, 다른 계절에 비해 여름철의 평균 농도 값이 감소하는 계절적 패턴은 잘 모의하고 있다. 또한, SO_2의 주요 배출원인 산업 시설에서의 화석연료 연소가 중국의 베이징, 허베이성, 장쑤성을 중심으로 북동부에 밀집되어 있기 때문에 YRD 지역의 북부에서 고농도가 집중적으로 발생하는 것을 확인할 수 있다.

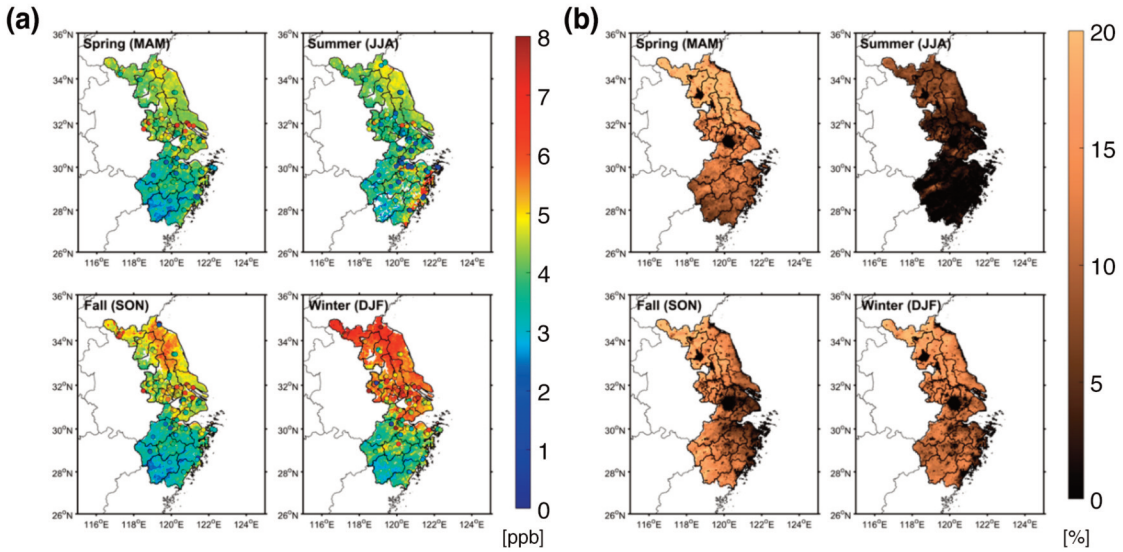

Fig. 7. (a) Spatial distribution of the ground-level SO_2 concentrations (ppb) by season in YRD. The background information represents the 2-step residual corrected RF model prediction, and the dots represent the station observations. (b) Spatial distribution of the temporal data coverage (%) for each pixel in YRD.

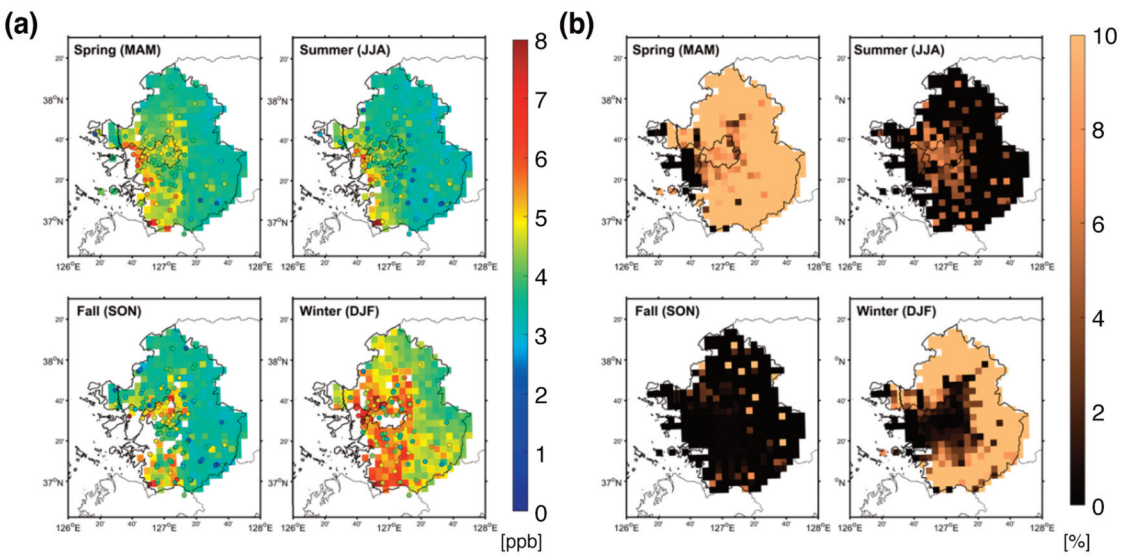

Fig. 8. (a) Spatial distribution of the ground-level SO_2 concentrations (ppb) by season in SMA. The background information represents the 2-step residual corrected random forest model prediction, and the dots represent the station observations. (b) Spatial distribution of the temporal data coverage(%) for each pixel in SMA.

Fig. 8은 한국의 SMA 지역에서의 계절별 SO_2 농도 변화와 결과 산출 가능 날짜의 비율을 함께 나타내고 있으며, SMA 지역에서 역시 모델 추정 값이 관측 값과 상당히 유사하게 모의하고 있는 것을 확인할 수 있다. 이 지역에서 역시 겨울에 고농도가 발생하고 여름철 농도가 낮아지는 일반적인 SO_2의 계절적 농도 변화 패턴을 나타내며, 2-step 잔차 보정 모델의 결과 분포 지도가 SMA 지역의 산업단지 조성 현황 분포와 매우 유사한 것을 확인하였다. SMA 지역에서 대부분의 산업시설은 황해안을 따라 파주, 김포, 화성, 인천, 평택 등의 SMA 서부 지역에 밀집되어 있으며 겨울철의 SO_2 지상농도 역시 서부를 중심으로 고농도가 발생하는 것을 확인할 수 있다. 반면, 서울에서 관측 값과 예측 값 사이의 상대적으로 큰 오차를 발견할 수 있는데, 이는 Fig. 8(b)에서 나타나듯 매우 높은 결측 비율에서 기인하는 것으로 보이며, 추후 입력 변수의 보간을 통한 보완 시 이러한 한계점을 더욱 개선할 수 있을 것으로 판단된다.

5. 결론

본 연구에서는 TROPOMI에서 산출되는 SO_2 연직 컬럼 농도 자료를 비롯한 다양한 위성 및 모델 자료를 융합 활용하여 기계학습 기반의 SO_2 지상 농도 추정 모델을 개발하였다. TROPOMI 위성 자료는 기존 연구에서 널리 활용되었던 OMI 위성 자료에 비하여 더 향상된 공간해상도로 정보를 제공하므로, 더 높은 공간해상도(6 km × 6 km)의 일별 SO_2 지상 농도 산출이 가능하였다. 개발된 모델은 일반화 성능 평가를 위하여 훈련에 사용되지 않은 위치 및 시간에 대한 관측소 값을 이용하여 평가되었다. 10-fold 교차 검증 결과는 본 연구에서 개발된 모델이 TROPOMI 연직 컬럼 농도 자료로부터 지상 농도를 추정할 수 있다는 것을 입증하였다. 2-step 잔차 보정 RF 모델의 PC 지표는 기울기 향상, R 향상, RMSE 및 rRMSE의 감소를 통하여 잔차 보정 과정이 유의미한 역할을 한다는 것을 보여주었으며, 모델 예측 값과 실측 값 간의 기울기 보정 효과가 가장 크게 나타났다. 계절 평균 SO_2 지상 농도 분포를 관측 값 분포와 비교한 결과 관측 값의 분포와 매우 유사한 분포를 보여주었으므로 SO_2 지상 농도의 공간적인 분포를 파악하는 데에 큰 기여를 할 수 있을 것으로 판단된다. 구름에 의한 오염 혹은 높은 알베도로 인해 발생하는 위성 자료의 결측을 보완하기 위해 최신 기법인 Generative Adversarial Networks (GAN), Convolutional Long Short-Term Memory (ConvLSTM) 등의 딥러닝 적용을 통한 입력자료 보간 활용 시 더 정확한 SO_2 지상 농도 표출이 가능할 것으로 보인다. 비록 TROPOMI 위성 산출물이 향상된 공간 해상도로 SO_2 연직 컬럼 농도를 제공할 수 있지만, 대류권 농도가 아닌 전층 연직 컬럼 농도 정보라는 한계점을 가지고 있다. 따라서, 대류권 연직 컬럼 농도 정보를 제공할 수 있는 타 위성(e.g. OMI) 산출물을 융합 활용한다면 더 향상된 결과를 산출할 수 있을 것으로 기대된다.

사사

본 논문은 환경부의 재원으로 국립환경과학원의 지원을 받아 수행하였고(NIER-2020-04-02-086), 과학 기술정보통신부 및 정보통신기획평가원의 대학ICT 연구센터지원사업의 연구결과로 수행되었음(IITP-2021-2018-0-01424).

References

Ahmad, M., K. Alam, S. Tariq, S. Anwar, J. Nasir, and M. Mansha, 2019. Estimating fine particulate concentration using a combined approach of linear regression and artificial neural network, *Atmospheric Environment*, 219: 117050.

Baawain, M.S. and A.S. Al-Serihi, 2013. Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network, *Aerosol and Air Quality Research,* 14(1): 124-134.

Bauduin, S., L. Clarisse, J. Hadji-Lazaro, N. Theys, C. Clerbaux, and P.-F. Coheur, 2016. Retrieval of near-surface sulfur dioxide (SO_2) concentrations at a global scale using IASI satellite observations,

Atmospheric Measurement Techniques, 9(2): 721-740.

Berman, J.D., P.N. Breysse, R.H. White, D.W. Waugh, and F.C. Curriero, 2015. Evaluating methods for spatial mapping: Applications for estimating ozone concentrations across the contiguous United States, *Environmental Technology & Innovation*, 3: 1-10.

Breiman, L., 2001. Random forests, *Machine Learning*, 45(1): 5-32.

Chen, J., J. Yin, L. Zang, T. Zhang, and M. Zhao, 2019. Stacking machine learning model for estimating hourly $PM_{2.5}$ in China based on Himawari 8 aerosol optical depth data, *Science of the Total Environment*, 697: 134021.

Choi, H., Y. Kang, J. Im, M. Shin, S. Park, and S.-M. Kim, 2020. Monitoring Ground-level SO_2 Concentrations Based on a Stacking Ensemble Approach Using Satellite Data and Numerical Models, *Korean Journal of Remote Sensing*, 36(5-3): 1053-1066 (in Korean with English abstract).

Choi, M., J. Kim, J. Lee, M. Kim, Y.-J. Park, B. Holben, T.F. Eck, Z. Li, and C.H. Song, 2018. GOCI Yonsei aerosol retrieval version 2 products: an improved algorithm and error analysis with uncertainty estimation from 5-year validation over East Asia, *Atmospheric Measurement Techniques*, 11(1): 385-408.

Combrink, J., R. Diab, F. Sokolic, and E. Brunke, 1995. Relationship between surface, free tropospheric and total column ozone in two contrasting areas in South Africa, *Atmospheric environment*, 29(6): 685-691.

Feng, L., Y. Li, Y. Wang, and Q. Du, 2020. Estimating hourly and continuous ground-level $PM_{2.5}$ concentrations using an ensemble learning algorithm: The ST-stacking model, *Atmospheric Environment*, 223: 117242.

Fernandes, A., M. Riffler, J. Ferreira, S. Wunderle, C. Borrego, and O. Tchepel, 2019. Spatial analysis of aerosol optical depth obtained by air quality modelling and SEVIRI satellite observations over Portugal, *Atmospheric Pollution Research*, 10(1): 234-243.

Fioletov, V., C. McLinden, N. Krotkov, M. Moran, and K. Yang, 2011. Estimation of SO_2 emissions using OMI retrievals, *Geophysical Research Letters*, 38(21).

Huang, G. and K. Sun, 2020. Non-negligible impacts of clean air regulations on the reduction of tropospheric NO_2 over East China during the COVID-19 pandemic observed by OMI and TROPOMI, *Science of the Total Environment*, 745: 141023.

Huang, K., Q. Xiao, X. Meng, G. Geng, Y. Wang, A. Lyapustin, D. Gu, and Y. Liu, 2018. Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain, *Environmental Pollution*, 242: 675-683.

Kharol, S.K., C.A. McLinden, C.E. Sioris, M.W. Shephard, V. Fioletov, A.V. Donkelaar, S. Philip, and R.V. Martin, 2017. OMI satellite observations of decadal changes in ground-level sulfur dioxide over North America, *Atmospheric Chemistry and Physics,* 17(9): 5921-5929.

Li, R., L. Cui, J. Liang, Y. Zhao, Z. Zhang, and H. Fu, 2020a. Estimating historical SO_2 level across the whole China during 1973-2014 using random forest model, *Chemosphere*, 247: 125839.

Li, T., Y. Wang, and Q. Yuan, 2020b. Remote Sensing Estimation of Regional NO_2 via Space-Time Neural Networks, *Remote Sensing*, 12(16): 2514.

Liaw, A. and M. Wiener, 2002. Classification and regression by randomForest, *R News*, 2(3): 18-22.

Lin, C.-A., Y.-C. Chen, C.-Y. Liu, W.-T. Chen, J.H. Seinfeld, and C.C.-K. Chou, 2019. Satellite-derived correlation of SO_2, NO_2, and aerosol optical depth with meteorological conditions over East Asia from 2005 to 2015, *Remote Sensing*, 11(15): 1738.

Qin, K., X. Han, D. Li, J. Xu, D. Loyola, Y. Xue, X. Zhou, D. Li, K. Zhang, and L. Yuan, 2020. Satellite-based estimation of surface NO_2 concentrations over east-central China: A comparison of POMINO and OMNO2d data, *Atmospheric Environment*, 224: 117322.

Seo, J., J. Yoon, G.-H. Choo, D.-R. Kim, and D.-W. Lee, 2020. Long-term Trend Analysis of NO_x and SO_x over in East Asia Using OMI Satellite Data and National Emission Inventories (2005-2015), *Korean Journal of Remote Sensing*, 36(2-1): 121-137 (in Korean with English abstract).

Shah, V., D.J. Jacob, K. Li, R.F. Silvern, S. Zhai, M. Liu, J. Lin, and Q. Zhang, 2020. Effect of changing NO_x lifetime on the seasonality and long-term trends of satellite-observed tropospheric NO_2 columns over China, *Atmospheric Chemistry and Physics*, 20(3): 1483-1495.

Shakeel, M., Q. Arshad, R. Saeed, T. Ahmed, H. Khan, M. Noreen, A. Ali, and A. Munir, 2015. Application of GIS in Visualization and Assessment of Ambient Air Quality for SO_2 and NO_x in Sheikhupura City, *Journal of Geography & Natural Disasters,* 5(150): 2167-0587.

Song, J., 2015. Bias corrections for Random Forest in regression using residual rotation, *Journal of the Korean Statistical Society*, 44: 321-326 (in Korean with English abstract).

Sulla-Menashe, D. and M.A. Friedl, 2018. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product, *USGS: Reston*, VA, USA, pp. 1-18.

Veefkind, J., I. Aben, K. McMullan, H. Förster, J. De Vries, G. Otter, J. Claas, H. Eskes, J. De Haan, and Q. Kleipool, 2012. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications, *Remote Sensing of Environment*, 120: 70-83.

Wang, X. and W. Sun, 2019. Meteorological parameters and gaseous pollutant concentrations as predictors of daily continuous $PM_{2.5}$ concentrations using deep neural network in Beijing-Tianjin-Hebei, China, *Atmospheric Environment*, 211: 128-137.

Wright, M.N., S. Wager, P. Probst, and M.M.N. Wright, 2018. ranger: A Fast Implementation of Random Forests, https://github.com/imbs-hl/ranger, Accessed on Jan. 10, 2020.

Xiao, Q., G. Geng, J. Cheng, F. Liang, R. Li, X. Meng, T. Xue, X. Huang, H. Kan, and Q. Zhang, 2021. Evaluation of gap-filling approaches in satellite-based daily $PM_{2.5}$ prediction models, *Atmospheric Environment*, 244: 117921.

Young, M.T., M.J. Bechle, P.D. Sampson, A.A. Szpiro, J.D. Marshall, L. Sheppard, and J.D. Kaufman, 2016. Satellite-based NO_2 and model validation in a national prediction model based on universal kriging and land-use regression, *Environmental Science & Technology*, 50(7): 3686-3694.

Yuan, Q., H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, and J. Wang, 2020. Deep learning in environmental remote sensing: Achievements and challenges, *Remote Sensing of Environment*, 241: 111716.

Zhan, Y., Y. Luo, X. Deng, M.L. Grieneisen, M. Zhang, and B. Di, 2018a. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment, *Environmental Pollution*, 233: 464-473.

Zhan, Y., Y. Luo, X. Deng, K. Zhang, M. Zhang, M. L. Grieneisen, and B. Di, 2018b. Satellite-based estimates of daily NO_2 exposure in China using hybrid random forest and spatiotemporal kriging model, *Environmental Science & Technology*, 52(7): 4180-4189.

Zhang, D., Y. Zhou, L. Zheng, R. Shi, and M. Chen, 2018. The spatial distribution characteristics and ground-level estimation of NO_2 and SO_2 over Huaihe River Basin and Shanghai based on satellite observations, *Proc. of 2018 International Society*

for Optics and Photonics, Remote Sensing and Modeling of Ecosystems for Sustainability XV, San Diego, CA, Sep. 18, vol. 10767, p. 107670L.

Zhang, G. and Y. Lu, 2012. Bias-corrected random forests in regression, *Journal of Applied Statistics*, 39(1): 151-160.

Zhang, Y. and W. Rossow, 1997. Estimating meridional energy transports by the atmospheric and oceanic general circulations using boundary fluxes, *Journal of Climate*, 10(9): 2358-2373.